National Climatic Data Center

DATA DOCUMENTATION

FOR

DATA SET 3721 (DSI-3721)

Gridded US Daily Precipitation
And Snowfall Time Series

March 24, 2003

National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001 USA

:

Table of Contents

1. **Abstract**: Two gridded data sets were produced for each element (currently for daily precipitation and snowfall for the period from January 1, 1948 to December 31, 1999, or 18,993 days). Directory US_gridded at the flux workstation contains the entire archive at the raid2 drive. The data are located in separate directories (SNOW and PRCP), each of which has two subdirectories (grid1 and grid2) that contain the output of the forth and sixth steps of the gridded procedure. Specifically, grid1 contains the time series only for A-cells (2752 and 2635 for precipitation and snowfall, respectively). Each A-cell that was included in the snowfall data set is also included in the A-cell group of the precipitation data set. The grid2 directory contains A and B cells daily time series (3341 and 3293 for precipitation and snowfall, respectively) that are interpolated (simulated) according to the algorithm of the sixth step and include the time series that are serially complete.

Gridded (with a 0.5° latitudinal and longitudinal resolution) daily precipitation and snowfall time series were compiled for the contiguous United States for the 1948-1999 period. The precipitation time series are serially complete and cover more than 96 percent of the land area of the lower 48 states and 100 percent of the country east of 105°W longitude. The snowfall time series cover about 95 percent of the land area of the lower 48 states. The nature of the data set is that for each cell the time series preserve the point precipitation (snowfall) distribution most typical for the cell and spatial correlation structure of the precipitation and snowfall fields. Extensive additional information about the origin and the time of the observation of each precipitation/snowfall datum is included into the data set. The data set was developed to meet the specific needs of those users whose studies require serially complete gridded fields of daily meteorological variables that preserve (resemble) as close as possible statistical structure of the station-based point fields (distribution and spatial correlation).

2. **Element Names and Definitions**: See Appendix a - Methodology

3. **Start Date**: 19480101

4. **Stop Date**: 19991231

5. **Coverage**: North America

   a. Southernmost Latitude: 25N
   b. Northernmost Latitude: 50N
   c. Westernmost Longitude: 125W
   d. Easternmost Longitude: 65W

6. **How to Order Data**:

   Ask NCDC's Climate Services about the cost of obtaining this data set.
   Phone: 828-271-4800
   FAX: 828-271-4876
   E-mail: NCDC.Orders@noaa.gov

7. **Archiving Data Center**:

   National Climatic Data Center
   Federal Building
   151 Patton Avenue
   Asheville, NC 28801-5001
   Phone: (828) 271-4800.

8. Technical Contact:

National Climatic Data Center
Federal Building
151 Patton Avenue
Asheville, NC 28801-5001
Phone: (828) 271-4800.

9. Known Uncorrected Problems: At some stations in early 1948 the absence of snowfall, precipitation, and snow on the ground was coded as missing values. They were simply skipped in the data set and during the digitization phase were interpreted as missing. In the example below this was found at the two Nebraska stations (Fullerton and Genoa 2 W). These stations were first and third in the preference list of four for use to represent the daily snowfall values in this cell in January-February. During the first five months of 1948, these were the only two stations available for this particular 0.5 x 0.5 grid cell that sporadically reported snowfall (see example below). As a result from 60 values for these two months in 1948, 9 values are reported as present and 51 others as missing. Among the nine present values only three are equal to zero and six others give non-zero values while the probability of the day without snowfall at this grid cell in these months is close to 0.9 (0.86 for January and 0.91 for February). This means that most probably all non-zero values of snowfall has been already reported for these two months in 1948. However, our infilling process at the sixth stage randomly added 6 more non-zero values instead of missing data. This cell is very well infilled with the daily snowfall data (99.2%) but the above mentioned difficulty has to be kept in mind in analyses of the 1948 snowfall and precipitation data.

10. Quality Statement: The data that come from the Current Data Rescue Data Set (prior to August 1948) are considered preliminary because they still are undergoing extensive quality control procedures. In our datum selection process this Data Set has a lowest priority and contributed only when all other four data sets had no datum at a given station at a given day. Generally, the user will be much "better off" using the 1948 data only as a supplement to the 1949 hydrological year (October 1948 to September 1949).

11. Essential Companion Datasets: None.

12. References:

Hughes, P.Y., E.H. Mason, T.R. Karl, and W.A. Brower, 1992: United States Historical Climatology Network Daily Temperature and Precipitation Data, ORNL/CDIAC-50, NDP-042, ESD Publ. No. 3778, Carbon Dioxide Information Data Center, Oak Ridge National Lab., Oak Ridge, TN. 54 pp. + Appendices. Updated to 1999.

National Climatic Data Center (NCDC), 1998: Surface Land Daily Cooperative. Summary of the Day. DSI-3200. Prepared by Lewis France. On-line documentation available from: http://www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html.

Frei, A., D.A. Robinson, and M.G. Hughes, 1999: North American snow extent: 1910-1994. *Internat. J. Climatology*, **19**, 1517-1534.

Kunkel, K., K. Andsager, and D.R. Easterling, 1999:Trends in heavy precipitation events over the continental United States. *J. Climate*, **12**, 2515-2527.

Appendix A - Methodology

INTRODUCTION

The idea of this approach belongs to Linda Mearns. The approach targets a specific group of users (ecology modelers) who need serially complete gridded fields of daily meteorological variables that preserve (resemble) as close as possible the statistical structure of the station-based point fields (distribution and spatial correlation). It is known that any interpolation and spatial averaging of the field affects these characteristics, and some of them (e.g., daily extremes) diverge significantly from the point values. Therefore, some of the steps of the process described below significantly deviate from standard gridding methodologies.

The fields that we are working with over the 48 states of the contiguous United States are:

- Mean daily precipitation and snowfall (24-hourly total),
- Mean daily snow depth (at the time of observation), and
- Minimum and Maximum surface air temperature

(as reported by the U.S. Cooperative observational network).

The Objective of the study is to reduce the station observations of these fields to a 0.5°x0.5° grid cell network. This reduction should preserve as close as possible the major statistical characteristics of the point field while at the same time characterize the typical climatic conditions of the appropriate grid cells.

The first step was to select a subset of the U.S. cooperative stations with long-term observations. The criterion used in this selection was the presence at least of 25 years of data during the 30-year reference period 1961-1990. This means that ≥83% of valid daily data should be present in the data set. This criterion was applied separately to each of the daily meteorological fields listed above for the contiguous U.S. and resulted in the selection of 58861, 53612, 4952[2], and 4164 stations for precipitation, snowfall, snow depth, and min/max temperature respectively. The stations that meet the above criteria were mapped and sorted by 0.5° x 0.5° grid cells. The results of this sorting are presented in Figures 1 and 2. These figures show that the eastern two-thirds of the contiguous U.S. can be easily covered by the station data with at least one station in the grid cell and/or with some sort of interpolation strictly from the adjacent cells. This is not always the case in mountainous and desert areas of the Western United States. Taking into account the small correlation radii of precipitation and snowfall fields, a decision was made to skip the following grid cells from the future analyses: those grid cells that do not have a single station inside them and (a) do not have at least one neighboring grid cell with the U.S. station data3 or (b) have an average elevation above 2,000

1 Originally, there were 5873 stations with precipitation records that meet this criteria but I added 13 stations that meet this criteria for snowfall and snow depth but for some reasons were slightly short of precipitation data for the reference 1961-1990 period.

2 Originally there were 5355 stations but I added six Maine stations that were slightly short of snowfall data to meet this criteria but were present in precipitation station subset. The reason for addition was to secure a complete coverage of eastern U.S. with gridded data at the last step of the procedure. The original 5355 stations left two 0.5° x 0.5° grid cells in Maine uncovered in this process. For snow depth we had originally 4917 stations that met our criteria but I added into the list additional 35 stations (19 in Maine and 16 in Northern Minnesota) for the same reason.

3 These cells are marked by "n" symbol in Figures 3 and 4.

meters4 or (c) are located primarily over ocean/sea/bay/lake5. Figures 3 and 4 illustrate the consequences of this decision for the daily precipitation and snowfall fields, respectively. In these maps, 0.5° x 0.5° grid cells are marked with 1 if the cell has at least one U.S. station with the data and are left blank if the cell does not have U.S. stations within it but the neighboring grid cells with the data allow interpolation into it. Other grid cells (including those outside the contiguous U.S.) are marked by "n," "X," and "O" characters.

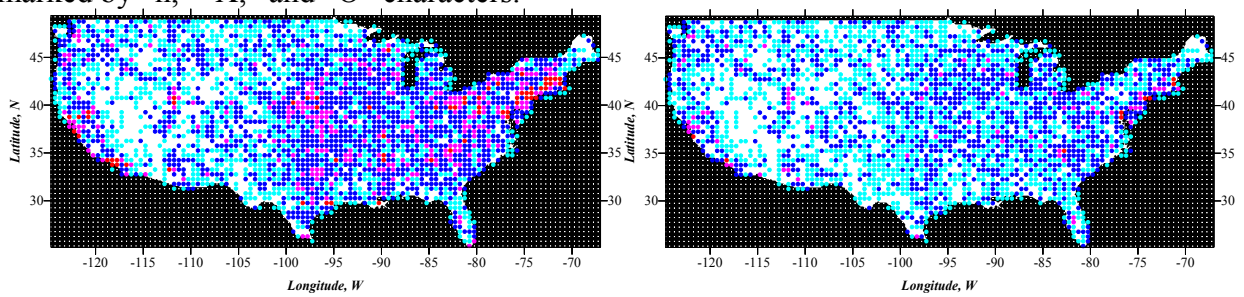


Figure 1. Map of the grid cells (0.5° x 0.5°) containing at least one cooperative station. Blue indicates a 0.5° x 0.5° grid cell with at least one station with a sufficient amount of daily data during the 1961-1990 period (i.e., more than 25 years of data); dark blue indicates grid cells with 2 to 3 stations; magenta indicates grid cells with 4 to 5 stations; and red indicates grid cells with 6 or more stations (maximum of 13). Left panel: precipitation; right panel: minimum/maximum temperature.

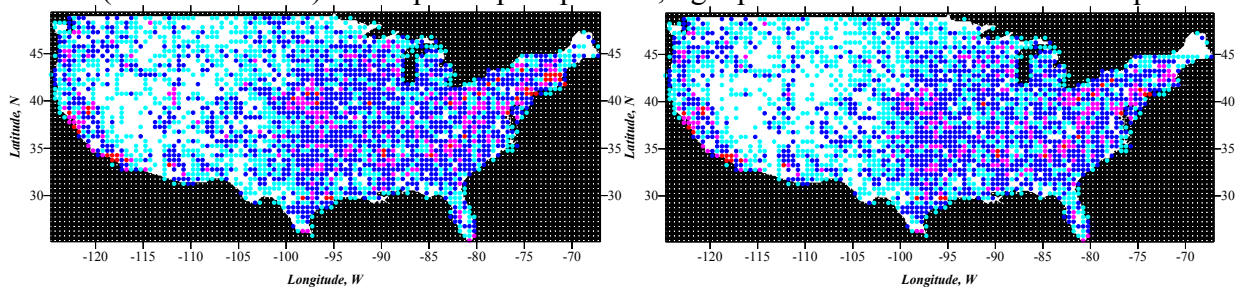


Figure 2. Map of the grid cells (0.5° x 0.5°) containing at least one cooperative station. Blue indicates a 0.5° x 0.5° grid cell with at least one station with a sufficient amount of daily data during the 1961-1990 period (i.e., more than 25 years of data); dark blue indicates grid cells with 2 to 3 stations; magenta indicates grid cells with 4 to 5 stations; and red indicates grid cells with 6 or more stations (maximum of 13). Left panel: snowfall; right panel: snow on the ground.

The processing of precipitation and snowfall fields is practically identical, and below I describe it using the example of daily precipitation. For snow depth some changes will be introduced to the final two steps of processing described below. The sequence of gridded fields to be generated was defined as snowfall, precipitation, maximum daily air temperature, snow depth, and minimum daily air temperature. Currently, only the two first two fields have been completed.

STEPS OF THE ALGORITHM

1. At the first step for each station and month, we estimate three parameters of a mixed gamma-distribution for precipitation:

4 These cells are marked by "X" symbol in Figures 3 and 4.
5 These cells are marked by "O" symbol in Figures 3 and 4.

$F(x) = P_{zero} + (1-P_{zero}) F_1(x,\eta,\lambda)$,

where P_{zero} is the probability of a day without precipitation, and $F_1(x,\eta,\lambda)$ is a two-parameter gamma-distribution function. η - is the shape parameter of this distribution, and λ is the scale parameter of this distribution. The mean value, μ, standard deviation, σ, and variation coefficient, C_v, of daily precipitation distribution can be expressed by the following formulas;

$\mu = (1- P_{zero})\eta/\lambda$,
$\sigma^2 = (1- P_{zero})^2\eta/\lambda^2$, and
$C_v = \sigma/\mu = \eta^{-0.5}$.

These parameters were estimated using all data available during the 50-year-long period 1949-1998 for each station and month. In some desert locations in the west in summer months this period was insufficient for the estimation of η and λ because of the extremely rare occurrence of precipitation. These values were marked as missing. For snowfall in the warm season, these parameters are also mostly undefined.

With these parameters at hand I performed several studies before the next step. First, I assured myself that they behaved reasonably compared to my previous findings for summer precipitation (Groisman et al. 1999 in *Climatic Change*) and common sense. Then I studied the dependence of these parameters from geographical coordinates and elevation using a simple linear regression inside 5° x 5° grid cell boxes. The results of these analyses support the previously found notion that the shape parameter of daily precipitation distribution is relatively stable geographically and changes little inside these boxes (especially compared to P_{zero} and λ). This strengthened my belief that during a short-distance interpolation inside of 0.5° x 0.5° cells (or maximum to the neighboring such cell), we can leave this parameter intact and change only the other two. This in turn gives a clear-cut solution to the issue of interpolation: if we know (find out from the station data inside the cell or by stochastic interpolation described in step 7) that there is a precipitation event inside the cell, then we should only adjust the scale (governed by the λ-parameter) of the available point measurements within the cell to make them "more representative" for the center of the cell location and for the cell mean elevation. On the contrary, if the most representative for the center of the cell site reports no precipitation event, then the issue of the precipitation intensity does not resurface at all.

Table 1. Mean grid box elevation (in meters) for 5 x 5 grid boxes over the contiguous United States and adjacent areas. The lowest line represents the 5° latitudinal band from 25°N to 30° N while the uppermost line the latitudinal band from 45° to 50°N. The first column represents the 5° longitudinal sector from 125°W to 120°W and the last column the sector from 70°W to 65°W.

Lat\Lon	125	120	115	110	105	100	95	90	85	80	75	70 W
45-50	521	996	1525	979	688	407	383	283	310	273	223	196
40-45	891	1520	1922	2018	1093	496	301	235	255	305	214	12
35-40	255	1514	1727	2312	1281	428	248	189	425	157	0	0
30-35	4	324	715	1637	998	287	75	118	115	4	0	0
25-30	0	5	59	1280	958	90	0	0	5	0	0	0

:
:

Number of stations with long-term precipitation time series located within each 5° x 5° grid box over the contiguous United States and adjacent areas.

Lat\Lon	125	120	115	110	105	100	95	90	85	80	75	70 W
45-50	113	92	97	84	107	106	81	57	13	0	3	14
40-45	137	78	118	88	136	233	226	191	162	210	275	14
35-40	146	67	81	140	138	285	212	217	273	217	11	0
30-35	5	107	87	78	115	234	193	210	203	31	0	0
25-30	0	0	0	0	11	104	17	1	66	0	0	0

Number of stations with long-term snowfall time series located within each 5° x 5° grid box over the contiguous United States and adjacent areas.

Lat\Lon	125	120	115	110	105	100	95	90	85	80	75	70 W
45-50	102	76	56	45	91	93	74	55	13	0	2	11
40-45	132	71	85	63	124	216	219	178	143	175	229	12
35-40	142	60	70	126	130	269	206	200	246	203	9	0
30-35	5	106	85	77	114	218	189	206	202	30	0	0
25-30	0	0	0	0	11	104	13	1	68	0	0	0

2. At the second step I calculated the regression equations for P_{zero} and λ as a function of latitude, longitude, and elevation inside each 5° x 5° grid box starting from –125°W and 25° N (low right cell corner) to –65°W and 50°N (upper left cell corner). In most of these boxes (inside the U.S. boundaries) there were plenty of stations (Table 1), and I calculated these regression equations for those boxes where at least 10 stations were present. For a few boxes (two southernmost boxes and one box in the northeast) with fewer than ten stations, the regression equations were later (at step 3) transported from the neighboring boxes. These regression equations were calculated as follows:

1. Correlations, R, between the climate parameter and each coordinate (lat, lon) and elevation were estimated. To characterize the elevation coordinate, I finally decided to use ln(100+[elevation in meters]) instead of the actual elevation for the λ-parameter and exp[elevation in km] for P_{zero}. Extensive experiments have shown that the shape of each of these relationships has some advantages compared to linear and others and describe the largest portion of the variance in regression equations. If max R_i^2 (i=1,2,3 corresponding to latitude, longitude, and elevation respectively) was below 0.05, I assumed that there was no sufficient spatial variability to account for. There were only a few such box-months. For example, for precipitation, there were 16 and 25 such boxes for λ and for P_{zero} respectively from the total of 550 5° x 5° box – months and 4 half-boxes in the Northwestern United States (see item 4 of this step below); for snowfall these numbers were 46 and 6 respectively and fore snow depth 22 and 1.

2. If $R_{min}^2 = \min R_i^2 \geq 0.05$, then the multiple regression to all three coordinates (i.e., x, y, and z) was calculated. Otherwise, the variable that corresponded to R_{min} was excluded and only the two remaining coordinates were used in the multiple regression. An algorithm that reduces a ***least absolute deviation*** (LAD) from the regression line instead of a least mean square (LMS) deviation was used. LAD is more robust than LMS; this can be important in dry regions/months

with P_{zero} well above 0.9 and, therefore, a small sample size, n, used to estimate the gamma distribution parameters [in our case, $n \approx 50$ years x 30 days x $(1-P_{zero})$].

3. After the LAD implementation, the goodness of fit, R_*^2, by the regression surface of the parameter field for each month and grid box was evaluated and compared with the appropriate set of R_i^2 (i=1,2,3). If max $R_i^2 > R_*^2$, then I selected the i-coordinate that delivers this maximum and used the LMS regression estimate of the parameter to this coordinate as the best approximation of spatial variability of this parameter in this grid box. In these situations, the contribution of two other coordinates was neglected (set to zero). For precipitation, there were 102 and 94 such box-months for P_{zero} and λ-parameters respectively from the total of 554 box-months.

4. The non-linearity of regression equations along the western coastal 5° x 5° grid boxes at step 2 is most prominently seen in two northeastern grid boxes west of 120°W and north of 40°N. These grid boxes are data rich and thus allow an additional partition. The non-linearity of the relationship here was addressed by the longitudinal division in two sub-boxes of each of these grid boxes at the step 2 along the Cascade Range.

As an output of this step for each 5° x 5° grid box with more than 10 stations within it (Table 1) for each month, two regression equations (with 4 or less non-zero coefficients) for P_{zero} and λ-parameters were constructed and stored in separate files. The two westernmost 5° x 5° grid boxes were additionally divided by two longitudinal sectors each along 121.25°W (the northernmost box that encompasses the Washington state) and along 122°W (box that encompasses most of Oregon). Additionally, I stored the information about those grid boxes/months, where/when all regression coefficients were zero (see item 1 of this step) and where/when effects of two (of three) coordinates were neglected (see item 2 of this step).

3. At the third step, for each 0.5° x 0.5° grid cell, A, over the contiguous United States marked by symbol "1" in Figure 3 and for each month, I constructed the list of the N "best" nearby stations for P_{zero} and λ separately (for these cells $1 \le N \le 13$), and sorted this list by the proximity to the cell center. Then I interpolated both parameters into the center of the cell. Afterwards, for each 0.5° x 0.5° grid cell, B, over the contiguous United States marked by blank symbol in Figure 3 and for each month, I selected the nearby 0.5° x 0.5° grid cells from the first subset (i.e., those A-cells marked by "1" in Figure 3), selected among them the "closest" grid cell, and interpolated the values of P_{zero} and λ into the B-cell from this A-cell. For each month and parameter, the "closest" A-cell can be different. The above procedure was done in several phases:

1. Initially, stations within each A - grid cell were sorted by their geographic distance from the cell center, the closest station to this center was the first. If for a given month, parameter, and 5° x 5° grid box, all three regression coefficients to coordinates were zero, then the value of the parameter (P_{zero} or λ) from this nearest station was assigned to the grid cell center.

2. If at least one regression coefficient for a given A-cell, month, and parameter was not zero, then the regression "distance" from the grid cell center was calculated for each station within the cell by formula:

   "distance" = $\partial P/\partial Lat.$[Lat. (Cell center) – Lat. (Station)] +
   $\partial P/\partial Lon.$[Lon.(Cell center) – Lon.(Station)] +
   $\partial P/\partial Elev.$[Elev.(Cell center) – Elev.(Station)],

   where $\partial P/\partial Lat$, $\partial P/\partial Lon.$, and $\partial P/\partial Elev.$ were the regression coefficients estimated at step 2 for parameter, P, by latitude, longitude, and the elevation function. Note that thus defined "distance" can be negative. The absolute value of this "distance" defines a new order of proximity from the grid cell center of each station in each month for each parameter. Therefore, the stations are sorted accordingly and new sequences of closest stations are selected and stored for each month, each parameter, and each A-cell. These regression "distances" are used in phase 3 of this step and in the next steps 4 and 6 to interpolate P_{zero} and λ into the grid cell center from the closest available station.

3. Using the smallest (by absolute value) regression "distances" defined in the previous phase of this step, I interpolated the parameter values from the closest stations into the grid cell center. It can be that for the same A-cell and the same parameter, the "closest" stations in different months were different. An additional restriction was implemented for the λ-parameter during this interpolation: the interpolated value could not be 100% more/less than λ in the closest station. This restriction was initially introduced to cover the sloppiness of the linear regression estimation at step 2 in the coastal grid boxes of the Washington state (the relationship there is non-linear across the Cascade Range during the cold season), and in the deserts in July (too small samples of precipitation events). After the additional partition of the northwestern grid boxes, this restriction for precipitation was applied only in July in 6 grid cells in the Mojave desert and for snowfall in 22 months for the A-grid cells (out of 31,560 cases) and only once in the B-cells along the southernmost edge of the seasonal snowfall appearance.

4. After interpolated values for each A-cell were created, the same procedure described in item 2 and 3 of this step was repeated for each B-cell. However, instead of "closest" stations available within each A-cell, for B-cells I used now the nearby A-cells ($1 \leq k \leq 4$) that have valid interpolated values of P_{zero} and/or λ- parameters.

As an output of this step for each 0.5° x 0.5° grid cell with at list 1 station within it (A-cell; Table 1) for each month, two files (for P_{zero} and λ) with the list of sorted nearest to the cell center stations within the cell together with the "distances" from this center were created and stored. For each 0.5° x 0.5° grid cell with at least one station within it or in the nearby cell (A- and B-cells in Table 1) for each month, interpolated values of P_{zero} and λ were constructed and stored. Additionally, I stored the information about those grid cell/months, where/when my interpolation procedure tried to change the λ-values more than twofold (see item 3 of this step).

4. At the forth step, the actual work with daily data started initially with the A-grid cells. For each of these cells, I retrieved all appropriate information acquired during the previous steps and pulled out the daily data for stations (up to 13 but no less than 1) within the cell and constructed two time series of gridded daily precipitation in this cell assigned to the center of this cell. The daily

information was collected from five archives of daily information: daily U.S. HCN6; U.S. Cooperative daily data set7 (the major source for the daily data since mid-1948); the U.S. daily data sets compiled by Dave Robinson8 for the northern half of the contiguous United States; by Ken Kunkel9 for Midwestern states and New Mexico; and at the National Climatic Data Center during the Data Rescue Effort. The last three data sets provide additional data coverage for the pre-1948 period and were used here mostly to infill the data for 1948 in our gridded data set. However, a provision was left to perform the entire gridding process starting from 1891 instead of 1948 as it was currently done. Specifically, the gridding for a specific date was done as follows:

1. All valid station data at this date were collected and sorted in the order of minimal "distance" for this month, cell, and λ. Then the closest station with the valid datum was selected, and this datum was assigned to the fist time series [unadjusted grid time series] and a scale corrected datum was assigned to the second time series [adjusted grid time series]. The ID of this station was also recorded. If all N stations within the grid cell had no valid data at this date, then missing code (-1) was inserted in both time series.

2. Scale correction was calculated as follows: The closest station in this month with valid data and valid λ (Λ; at few desert stations in summer months this parameter was not estimated due to short samples of the rain events) has a "distance," D, from the grid cell center estimated for λ for a given month. Therefore, a best available estimate of λ interpolation from this station to the grid cell center is $D+\Lambda$, and the scale correction to the precipitation value at this station to better reproduce the value in the grid cell center is $\Lambda/(D+\Lambda)$. As at the previous step, we trimmed the estimate of this scale correction by not allowing it to leave the closed interval [0.5; 2]. This scale adjustment obviously did not affect zero precipitation values.

As an output of this step, we generated for each A-grid cell (a total of 2752, 2635, and 2523 cells for precipitation, snowfall, and snow depth, respectively) files that contain two daily time series of gridded precipitation/snowfall value estimates for the 1948-1999 period (first of the closest station with valid datum and the second the same datum but scale-adjusted). It must be noted that in different dates the values of different stations can constitute this datum. Moreover, even if all stations within the grid cell did have data during the entire period in question, the different months may have had values from different stations because the order of the absolute values of "distances" from these stations to the grid cell center could be different. Therefore, in addition to each datum, the actual ID (six-digit cooperative number) of the current "closest" station is stored in the next column adjacent to the daily datum, and the actual time of precipitation observation for each of this datum (see more in the aftermath comments on this aspect of our daily data). Additionally, I stored

---

6 Hughes, P.Y., E.H. Mason, T.R. Karl, and W.A. Brower, 1992: United States Historical Climatology Network Daily Temperature and Precipitation Data, ORNL/CDIAC-50, NDP-042, ESD Publ. No. 3778, Carbon Dioxide Information Data Center, Oak Ridge National Lab., Oak Ridge, TN. 54 pp. + Appendices. Updated to 1999.

7 National Climatic Data Center (NCDC), 1998: Surface Land Daily Cooperative. Summary of the Day. TD3200. Prepared by Lewis France. On-line documentation available from: http://www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html.

8 Frei, A., D.A. Robinson, and M.G. Hughes, 1999: North American snow extent: 1910-1994. *Internat. J. Climatology*, **19**, 1517-1534.

9 Kunkel, K., K. Andsager, and D.R. Easterling, 1999:Trends in heavy precipitation events over the continental United States. *J. Climate*, **12**, 2515-2527.

for each grid cell the information about station order allocation and "distances" for each month and the percent of the data coverage of the gridded time series during the 1948-1999 period. Usually, this percent is close to 100%, but there are grid cells that have only about 60-70% of daily data.

5. At the fifth step, the preparation work for infilling the missing data in the gridded daily precipitation time series was done. The step was identical for precipitation and snowfall and thus is described for precipitation only. Specifically, for each A- and B-grid cell I calculated the conditional probability of the precipitation event absence within this cell, if the nearby cells had or did not have precipitation, or the information about this cell was absent in any combination. There are a maximum of four nearby cells for each grid cell and three situations (no information, no precipitation, and precipitation) each day. Thus, for each cell, I calculated $3^4 = 81$ conditional probabilities. The time series generated at the forth step for A-cells were used as an input, and these conditional probabilities were estimated empirically for each A-cell and month. Then, for each B-cell the neighboring A-cells were defined (there is at least one such cell), and the conditional probabilities for B-cell were defined as the average of available estimates from these neighboring cells. Not all combinations happen during the 52 years used in this estimation, and some of the estimates of these probabilities are absent (marked as missing). The conditional probabilities in these neighboring cells have a peculiarity: at these cells, the combinations that include the precipitation events (or the "no precipitation " events) in the adjacent B-cell were always absent, because the B-cell was empty throughout the entire period of observations. Therefore, each of these combinations was replaced by the appropriate combination that include the "no information" event in the adjacent B-cell. This replacement increased the countrywide non-trivial conditional probability estimates10 used at the next step from 35% of all situations with missing values to 50%. The output of this step is used in the final sixth step.

6. At the sixth step, the infilling of missing values of the time series generated at the forth step for each A-cell was performed and, additionally, the artificial daily precipitation time series were generated for each B-cell. A random number generator and previously estimated conditional probabilities (step 5) and parameters of the gamma distribution (in each A- and B-grid cell (step 3) are used at this step.

1. For each A-cell, I read the gridded time series generated at the forth step and the time series from neighboring A-cells (if any). When I found a missing value in the second (i.e., scale adjusted) time series in this cell, I investigated the neighboring cells for this date, selected the appropriate combination of events around the cell, and retrieved the empirical conditional probability, P'_{zero}, for this combination defined at the fifth step. If this probability was missing, then the P_{zero} estimate interpolated into the grid cell center was taken.

2. A random number from [0,1) was selected and compared with P'_{zero}. If it was less then P'_{zero}, then a "no precipitation" event at this day was assumed and zero was infilled instead of missing value. Otherwise, we assumed that there was a non-zero precipitation event, and its intensity was estimated with the help of the another random number. This number was generated and converted into a value sampled from a two-parameter gamma-distributed random variable. These

10 Non-trivial P'_{zero} at a given date means that it is somewhat different from the climatological value of $P_{zero.}$ and was defined using the current situation at the neighboring stations.

two parameters were (of course) those interpolated into the cell center at step 3. This value replaces the missing datum.

3. The same procedure was performed for each B-cell, except all values in it were initially missing.

4. Total number of A-and B grid cells for precipitation is 3341 (589 of them are B-cells). The portion of infilled values in both A- and B-cells for precipitation was 21% (100% in B-cells and 4% in A-cells; most of infilling in the A-cells was in the first half of 1948, when our major source of the data had 5 times less information than in the end of the year). In half of the days with missing values (16%, if A-cells are considered separately), completely random values were generated instead of missing datum, when P_{zero} was used instead of P'_{zero}. All other infillings were "improved" and accounted for the presence of additional information available from neighboring grid cells. For snowfall, in both A- and B-cells (totally 3293 and 658 of them are B-cells) 24% (5% values in the A-cells) were initially missing. In half of the days with missing values (15%, if A-cells are considered separately), these missing values were infilled with completely random values with the appropriate climatological distribution. The reason for a better performance of the infilling process for A-cells compared to the B-cells was in their locations. A-cells were mostly surrounded by the A-cells too. Thus the conditional probabilities, P'_{zero}, calculated at the previous step were usually at hand for the use in the infilling process for A-cells. B-cells were located in clusters in the western third of the country and the presence of these clusters restricted our ability to empirically estimate P'_{zero} -values. Therefore, for B-cells, we had to use climatological values of P_{zero} for infilling in ~50% of situations.

5. Only in a few cases (in 125 days for the entire 1948-1999 period for all 3341 A- and B-cells) over the contiguous U.S., we were not able initially to generate even random values of non-zero daily precipitation due to missing λ-parameters for the grid cell. This happens in the regions and months where precipitation is so rare that it is impossible to estimate parameters of the gamma distribution and, thus, when the random generator created the non-zero precipitation event, we could not simulate its intensity. For precipitation in the contiguous United States, these are quite rare situations (125 from total of 52 millions) but for snowfall this could be a problem. Therefore, the last step of the infilling process for these situations was modified and the parameters of previous or next months, whichever are closer, to the month without the λ- and η-estimates were used. This allows to cover the "last rainfall/snowfall before the dry/warm season starts" and or the "first rainfall/snowfall after the dry/warm season ends" events.

The output of this final step was 3341 precipitation (and 3293 snowfall) files, one for each A- and B-cell, with three time series in it: the first two were those constructed at step 4 (with descriptive information for A-cells only; and with all missing codes for B-cells), and the third contained the serially complete time series of daily precipitation/snowfall assigned to the grid cell center.

AFTERMATH COMMENTS

1. Times of observation.

We have the digital history of the time of precipitation and temperature observations at 7019 U.S. cooperative stations from 1951 to date11. For precipitation, the overlap of this station list with our 5886 station list was close to 99%. In other words, 59 precipitation and three temperature stations did not originally have digital time of observation histories. Therefore, I manually infilled the entire history of the times of observations for these stations. Times of observations have changed dramatically and systematically during the post-WWII period,[11] and future users may like to know and use this information in their analyses of our data set. The two-symbol codes used for these times of observations are presented in Table 2. Digital history of observations provides year and month of the change of the observation time but not the day. Therefore, during the processing and the fourth step, the middle of the month was considered as a breaking point when the metadata reported a change in observation times.

Table 2. Codes used to characterize times of observation of precipitation and temperature

Two symbol code	What does it mean
From 01 to 24	local standard times from 1 AM to 12 midnight
SR	Sunrise
SS	Sunset
PM	Sometimes in P.M.
VR	variable
AR	after rain (variable)
30	monthly/irregular (variable)
77	Variable
88	Irregular
99	Missing

2. Snowfall measurement correction.

I did not introduce the bias corrections into precipitation measurements. For rainfall, these corrections are small and can be assumed to be equal to factor of 1.04 but for frozen precipitation the relative biases are large (up to 60% in the northern Great Plains for open site locations). Unfortunately, the accurate introduction of these corrections requires a lot of metadata (gauge type and exposure) and supplementary information (wind and precipitation type among the most important). Gene Peck and I are slowly advancing state by state with this work but there is still a long way to go. Therefore, I plan later to leapfrog the problem in our gridded data set using the snowfall and temperature information and formal relationship between three of them and the water equivalent of freshly fallen snow under a specific temperatures at given sites/cells. This raises an issue of the coordinated gridded data sets that have the temperature, snowfall, precipitation and snow depth values (adjusted to the grid cell center) in each given day originated from the same station. This is the best way to secure the consistency among the fields.

3. Coordinated data set.

---

11 Groisman, P.Ya., B.-M. Sun, and R.R. Heim, Jr., 2000:Ttrends in spring snow cover retreat over the U.S. and the effect of observation time bias. Proc. of the Eleventh Symposium On Global Change Studies, Long Beach, California, 9-14 January, 2000. Amer. Meteorol. Soc., Boston, Mass.,54-57.

To address the previous comment, I can modify the future processing of the other variables in such a way that the priorities (defined by "distances" at the third step of the above processing) will be "frozen" and taken from those stored for the precipitation processing.

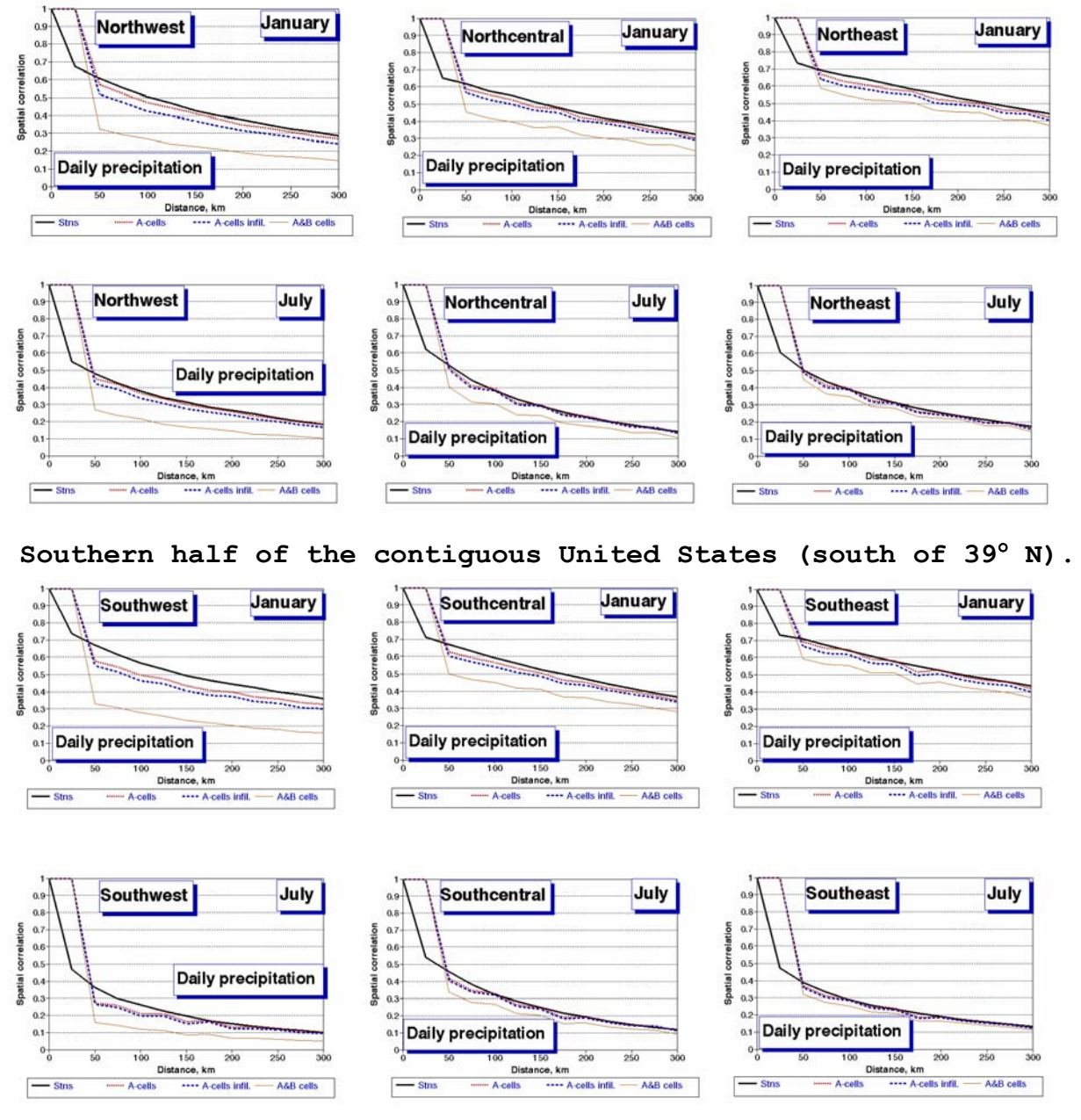
4. **Snow depth peculiarities.**
   This variable has memory that allows a better infilling process. At step 6 I used the past/next day information to construct the estimates of conditional probability of the presence of snow on the ground and maximum temperature data. Moreover, we may like to use the coordinated gridded snowfall and temperature data sets to estimate the infilled amount of the snow on the ground in the last step of the infilling process. The routine that unfolds this process in under construction and will follow that for T_{max}.

5. **Northward expansion.**
   It is not feasible to extrapolate these data southward across the Mexican border due to the low density of the Mexican station network available at NCDC. But, it is possible and very promising to do this in the northward direction using the dense high quality Canadian network in Southern Canada (south of 55°N). Collaboration with the Canadian AES specialists is warranted for this endeavor and snowfall is the first (easiest) candidate for this effort.

6. **Spatial correlation fields.**
   Figures 5 and 6 present examples of the comparison of spatial correlation functions for daily precipitation (figure 5) and snowfall (figure 6) respectively for six regions of the contiguous United States. Three of them were selected north of 39° N (northeastern, east of 90° W; northcentral (west of 90° W but east of 105° W; and northwestern, west of 105° W) and three south of 39° N with the same longitudes as dividers for southeastern, southcentral, and southwestern regions. Spatial correlation functions were estimated for the station values, for the "original" A grid cell values only, for original and infilled A grid cell values, and for all original and infilled A and B grid cell values. Calculations were made under the isotropic assumption, which means that, while for each pair of stations/gridcells the spatial correlation was calculated separately, these correlations were then grouped within the region as functions of the distance between the stations/gridcells. Thus, all directions in this estimation were considered equivalent or isotropic. The first "distance step" that was used for spatial correlation is [0 to 25 km] and is valid only for stations, because the distance between the centers of 0.5° x 0.5° grid cells is larger than 25 km.

   Analysis of Figures 5 and 6 shows that both the original only and infilled A-grid cell precipitation and snowfall data reasonably well preserve the spatial correlation function with exception of the nearest 25-50 km, where the distortion of this function is inevitable due to the nature of the gridding process. In the western U.S. (west of 105°W) where most of the B-cells are located, our infilling process degenerates significantly the spatial statistical structure of daily precipitation and snowfall. In this part of the contiguous U.S., B-cells are located in clusters and thus our attempts to use the neighboring cells to estimate the conditional probability of precipitation events (on step 5) more often remain unsuccessful and thus, at step 6, these cells are infilled with the values that are spatially incorrellated.

Northern half of the contiguous United States (north of 39° N).



Southern half of the contiguous United States (south of 39° N).

Figure 5. Spatial correlation function of daily precipitation over the contiguous United States (isotropic approximation). Station data fields are compared versus three types of gridded fields: A-cells only, infilled A-cells, and infilled A- and B- cells. Longitudinal partitions of the regions are along 105° W and 90° W.

Northern half of the contiguous United States (north of 39° N).

:

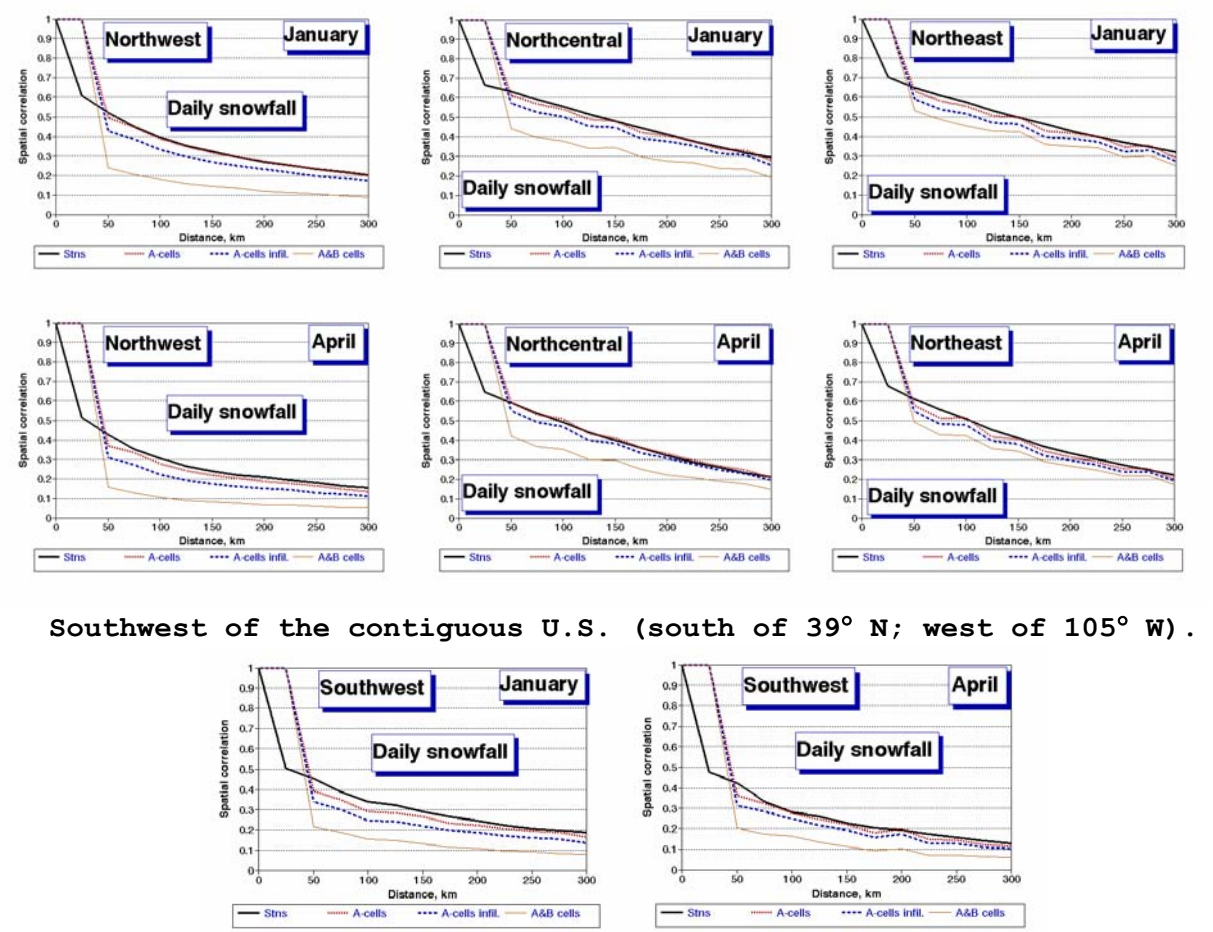


Figure 6. Spatial correlation function of daily snowfall over the contiguous United States (isotropic approximation). Station data fields are compared versus three types of gridded fields: A-cells only, infilled A-cells, and infilled A- and B- cells. Longitudinal partitions of the regions are along 105° W and 90° W.

PRESENT STATUS

Currently, two variables, snowfall and precipitation, are ready for use, and the snowfall archive is final (we do not need to adjust the linear measure of snow rulers). The snow depth variable has also been processed up to the fourth step (i.e., ".grid1" files for A-cells have been generated). The next step will be a generation of the T_{max} gridded fields. Then we shall compose the interpolated snow depths in the B-cells. The T_{min} field, due to a complexity of the task, will be the last in the queue. For this field it is our intention to use as much information as possible, including synoptic data, in order to distinguish advective and radiative factors that contribute to this field's spatial variability.

We plan to release these data to the public via the NCDC web site, and the necessary preparations for the release of the snowfall and precipitation archives are currently on the way.

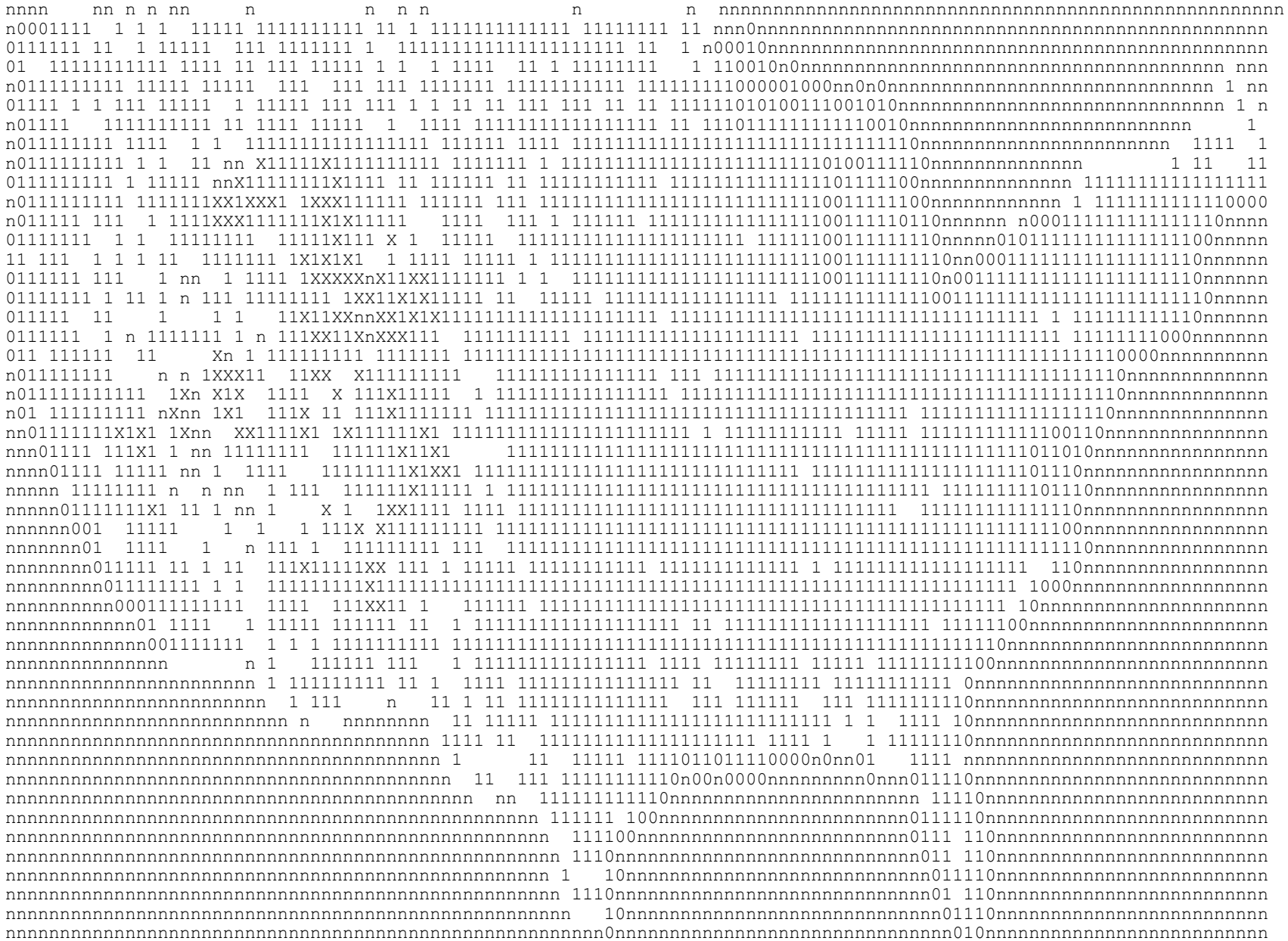


Figure 3. 0.5° x 0.5° grid cells over the contiguous U.S. Precipitation data availability described in text.

Figure 4. 0.5° x 0.5° grid cells over the contiguous U.S. Snowfall data vailability described in text.

:
: 21: